August 1975

PROGRESS IN SPEECH UNDERSTANDING RESEARCH AT SRI

Donald E. Walker

PROGRESS IN SPEECH UNDERSTANDING RESEARCH AT SRI

SUMMARY

This paper describes the current status of research being performed by Stanford Research Institute on the development of a speech understanding system capable of engaging a human operator in a conversation about a specific task domain. The design and implementation of the system are being carried out in conjunction with the System Development Corporation. Following a brief overview, the following system components are discussed in turn: System Control, Language Definition, Semantic Analysis, and Discourse Analysis and Pragmatics.

A.   Introduction

   1.   Project Objectives

   This paper describes the current status of research being performed by Stanford Research Institute (SRI) on the development of a speech understanding system capable of engaging a human operator in a conversation about a specific task domain.[1] This project is part of a five-year program of research sponsored by the Information Processing Techniques Office of the Defense Advanced Research Projects Agency.[2]

   The long term objective of the research at SRI on speech understanding is to develop the technology that will allow speech understanding systems to be designed and implemented for a wide variety of different task domains and environmental constraints. Early in 1974, SRI began to work cooperatively with the System Development Corporation (SDC) on the design and implementation of

------

a joint system. The first major step toward the SRI long term
objective is completion with SDC of this system in substantial
satisfaction of the specifications presented in the Newell Report
(Newell et al., 1973).   We expect to complete by fall 1975 a
'milestone system' that will have most of the components required
for the 'five-year' system.   This paper summarizes the
contributions that have been made by SRI to the development of
this milestone system. A detailed description is provided in the
most recent Annual Report of the project (Walker et al., 1975).

    2.   Background

        For three and a half years, SRI has been participating
with other ARPA/IPTO contractors in a major program of research on
the analysis of continuous speech by computer. During the first
year of the SRI project, the domain chosen provided interactions
with a simulated robot that knew about and could manipulate
various kinds of blocks.[3] The system implemented during this
period made major use of procedures developed by Winograd (1971)
for understanding sentences in natural language entered as text.

        During the second year of the project, a new task domain
was chosen: the assembly and repair of small appliances. This
change was made to provide for more complex interactions of a user
with the system, entailing a sequence of goal-directed subtasks.
------
[3] For descriptions of these initial efforts, see the First
Annual Report for the Project (Walker, 1973a), Walker (1973b), and
Paxton and Robinson (1973).

Major modifications were made in all parts of the system, the most important of which was the development of a new parsing strategy.[4]

SRI began collaborating with SDC on the development of a system following the Midterm Evaluation of the total ARPA Speech Understanding Research Program. Because of the similarity of the design concepts for the two contractors, it has been possible to combine features and components of the two most recent systems of each in building the new system architecture.[5] Work on signal processing, acoustics, phonetics, and phonology at SDC is being coordinated with work on parsing, syntax, semantics, pragmatics, and discourse analysis at SRI. There is shared responsibility for system design, for the specification of task domains, and for work on prosodics.

Two task domains have been selected for the duration of the current five-year program:

(1) Data management of a file containing information about selected ships from the fleets of the United States, the Soviet Union, and the United Kingdom.

(2) Maintenance of electromechanical equipment in a

------

[4] See the Second Annual Report for the project (Walker, 1974a); also see Walker (1974b), Paxton (1974), Becker and Poza (1975), Deutsch (1974), and Robinson (1975). Walker (1973c) provides a perspective on the transition from the first to the second versions of the system.

[5] For an overview of the previous SDC efforts and references to other SDC papers, see Ritea (1974).

workstation environment with the system   as   a   computer
consultant.

Since we began working with SDC,   most   of   our   activities   have
concentrated   on   the   first   domain,   but a substantial amount of
effort has   gone   into   ensuring   the   generality   of   our   system
structure and its appropriateness to the second domain.

The task domains selected are significantly different in
kind.   Together,   they represent the two major kinds of knowledge
identified in artificial intelligence   research: state   knowledge
and process knowledge.   State knowledge captures information about
a static world, all the facts that hold at a particular instant in
time   or   for   all   time.   Retrieving information from a formatted
file is a   representative   task   over   state   knowledge.   Process
knowledge embodies a dynamic model of the interrelations among the
elements of a world so   that   change   over   time   can   be   handled
directly.   Repairing   an   air   compressor or an automobile engine
exemplifies a relevant task.

The work on the second task domain   is   complemented   by
the   activities of a companion project at SRI that is developing a
comprehensive 'computer based consultant' (CBC)   system.[6]   That
system   is   designed   to   guide a technician in the maintenance of

------
[6] ARPA Contract No.   DAHC04-75-C-0005, SRI   Project   3805.   See
Nilsson   et   al. (1975) for the most recent Annual Report and Hart
(1975) for an overview of the project.

electromechanical equipment in a workstation environment. Our speech understanding system can provide the basis for communication with the computer in natural language.

B.   Overview of the System

1.   Introduction

An initial version of the cooperatively developed speech understanding system has been implemented and tested at SDC. The acoustic processing is provided by the Raytheon 704, and the rest of the system is programmed in SDC/LISP on the IBM 370/145. In addition, the parser and the syntactic, semantic, and discourse components have been exercised extensively at SRI on the PDP-10, with simulations of the acoustic, phonetic, and phonological components. These versions of the higher level language components are programmed in INTERLISP. More extensive testing of the total system will be conducted when INTERLISP/370 is available on the IBM 370/145 and when other components are reprogrammed for that computer in CRISP, a new programming system now under development at SDC. For the milestone system, SDC will replace the Raytheon 704 with an acoustic preprocessor consisting of a PDP-11/40 and an SPS-41 special purpose digital signal processor.

The following summary provides a perspective on the distinctive characteristics of the SRI contributions to the current system. The system control, embedded in the parser, focuses the operation of the entire system to minimize both

storage   requirements   and   the   time   spent   on   incorrect
interpretations.   A   language   definition   system   provides   a   means
for   integrating   the   various   sources   of   knowledge   in   the   system.
The   language   definition   itself,   based   on   studies   of   protocols
gathered   from   actual   performances   in   task-oriented   dialogs,
includes   information   from   acoustics,   phonetics,   phonology,
prosodics,   syntax,   semantics,   pragmatics,   and   discourse.   A   new
semantic   network   representation,   which   partitions   the   net   into
spaces,   has   proved   particularly   well   suited   for   working   with   the
two   task   domains.   Discourse   procedures,   building   on   the
semantics,   establish   a   discourse   history   so   that   information   from
previous   utterances   (and,   ultimately,   from   the   task   environment)
can   be   used   in   the   analysis   of   the   current   utterance.
Descriptions   of   these   developments   and   of   the   work   in   progress   are
presented   in   the   rest   of   this   paper.   These   presentations   will
serve   as   an   introduction.   Details   on   the   various   system
components   are   contained   in   the   current   Annual   Report   (Walker   et
al., 1975).

        2.   Status of the System Components

        a.   System Control

                The parsing system   coordinates   and   controls   the
other   system   components   in   the   process   of   understanding   an
utterance.   A   computationally   efficient   internal   representation   of
the   various   knowledge   sources   is   established   through   the   language
definition   system,   providing   a   uniform   way   of   integrating

different kinds of information. The external representation of the language definition is described under item 3 below. Words and phrases can be predicted on the basis of context, and phrases can be built up from words that have been identified acoustically in the utterance.

During the search for a complete interpretation of the utterance, a complex data structure called a 'parse net' is built up. The various tasks corresponding to alternative analyses are assigned priorities and scheduled according both to their estimated value and to a focus of activity that takes into consideration processing time and current storage requirements. When the performance of a task results in the prediction of a word at a specified place in the utterance being processed, various alternative phonological forms of that word are mapped onto the acoustic data for that place, and a score denoting the degree of correspondence is returned. Subsequently, when a phrase containing that word is predicted, another mapping is done to take into account coarticulation effects of the words on each other. The parser stops and calls a response function when it has an interpretation for the entire utterance or when it reaches a specifiable limit either on the number of tasks to be performed, on the lowest value of a priority it will accept, or on the amount of space it can use.

Efficiency has been a major motivating factor in the design of the parsing and language definition systems, with

respect both to the effort required by the people who are entering data and to the actual computations carried out inside the computer. A language definition compiler automatically converts rules as a linguist would write them into a form optimal for machine processing. The parse net brings together work on common substructures to eliminate duplication of effort. In addition, the various ways in which the same information can be used in different internal operations are anticipated, and, for computational efficiency, separate representations are constructed that are optimal for each use.

  b.   Language Definition

      The subset of natural spoken English that the system is designed to understand is specified by the language definition (LD). This component in a question-answering system is usually called a 'grammar', but our LD takes into account such a variety of linguistic information that 'grammar' does not adequately encompass it. The LD has two major parts:

  (1) A collection of basic units, called 'word definitions' (WD), which correspond roughly to words and together form a lexicon.

  (2) A collection of definitions of rules, called 'composition rule definitions' (CRD), for combining words and phrases into larger units.

Each CRD contains statements that assign attributes to the

resulting   unit   based   on   available   acoustic,   phonetic,
phonological, prosodic, syntactic,   semantic,   pragmatic,   or
discourse information. A CRD also contains factor statements that
establish how well the resulting unit fits the corresponding part
of the utterance, on the basis of all the determinable attributes.

Since October 1974,   the   language   definition   has
been   extended,   as   well as refined, to adapt it to the discourse
found in protocols collected for the data management task domain
during   the   summer   and   fall.   (Before that time, it defined a
language we assumed would be relevant for querying a small data
base   drawn   from   Jane's   Fighting   Ships.) New definitions were
added   for   elliptical   utterances   and   for   limited   comparative
expressions   involving   numbers.   Pragmatic factors were added to
existing definitions to adapt the LD to the   high   frequency   of
WH-interrogatives   and   elliptical   nominals.   By the end of 1974
more than 60 phrase types and 30 syntactic categories had been
defined,   and   the   LD   had   been   tested extensively on text and
simulated acoustic input and in a limited fashion on actual
acoustic input.

Further extensions to the language definition are
being   made   on the basis of analyses of additional protocols from
both task domains. CRDs are being written for additional phrase
types that are typical of the discourse required for the tasks and
sufficiently tractable to be put into the system and tested in a
reasonable time.   These include definitions for some kinds of

quantification, limited coordination, relative clauses, and compound nominals.   They will be ready for testing by the end of the current contract.

Earlier this year, SRI and SDC, together with the Speech Communication Research Laboratory (SCRL), established a set of conventions for transcribing protocols from our task domains, marking pauses (both silent and 'filled'), tonic syllables, and pitch direction. The data from these transcriptions are being used to revise the prosodic statements currently in the LD.

Further work on prosodics will be based on our judgment that the acoustic phenomena promising the most immediate returns for a prosodic component are silence and duration. The matrix of acoustic and phonetic data for one of the early submarine protocols was handmarked to locate pauses and to identify word durations. A concordance was compiled that brought together, in context, all occurrences of each word and pause, in order of increasing duration. These data allowed us to make comparisons and form hypotheses regarding the distribution of pauses and, in particular, the correlation of pauses with word boundaries and with word durations. We are arranging to test these hypotheses on the next round of protocols from different speakers. Our first comparisons support observations reported in the general literature on prosodics, which indicate that it should be possible to specify minimal durations for some kinds of words (stressed 'content' words) and conditions on lengthening of

unstressed words before pause.

We plan to use other acoustic attributes of words to distinguish among a set of words that are predicted for a particular place in the utterance. A preliminary scheme for classifying words on the basis of strong acoustic clues in their initial and final syllables has been developed. It is now being implemented at SDC and will be tested during the summer. Simulated tests on text with and without this lexical subsetting capability lead us to expect a significant improvement in parsing efficiency. Adding prosodic cues to the procedure should increase its discriminatory power.

c.   Semantic Analysis

The semantic component that has been developed for our speech understanding system consists of two major parts:

(1) A semantic network coding a model of the task domain.

(2) A battery of semantic composition routines that are directly coordinated with the language definition to build network representations of utterances.

Our semantic nets differ from other network representations in that the nodes and arcs of our nets are partitioned into units called spaces. These spaces group information into bundles that help to condense and organize the semantic knowledge base of the system. Specifically, partitioning facilitates quantification,

which in turn makes possible the description of generalized categories of objects, situations, and events. The organization of knowledge in terms of hierarchies of categories results in a more economical storage of information with properties common to all elements of each category being stored only once at the category level. (It remains clear that these properties are properties of the category members and not of the category itself.)

Net partitioning also provides a uniform mechanism for distinguishing hypothetical and imaginary situations from reality, a property of considerable importance in dealing with dynamic domains (such as our computer consultant task) characterized by multiplicities of alternative future states. The semantic composition routines that form a part of each language definition rule call on the information in the network to help understand the meaning of each phrase. Outputs from these routines are network fragments whose structures follow the same encoding conventions followed in the encoding of knowledge in the rest of the network.

We are currently experimenting with an improved set of network manipulation functions that are more efficient than their predecessors and that allow the network to be divided up in multiple ways. One of the new network groupings is being used to establish contexts (and hierarchies of contexts) within the net for use in discourse analysis. The revised network functions also

are being integrated into the semantic composition routines in a
way that will eliminate both the need for the 'intermediate
language' used in our previous work and the need to copy portions
of the network in cases of ambiguity or uncertainty.

       While modifying the semantic composition routines
to use the revised network manipulation functions, several other
improvements are being made as well.   Our present system uses
sequences of code especially written for each verb to associate
surface cases with deep semantic cases. These code sequences are
being replaced by a two-way case mapper that will interpret a
brief statement of case information included with the entry of
each verb-like member of the exicon to map from surface into deep
cases and vice versa. The added ability to map from deep to
surface cases will facilitate semantic prediction and the
generation of answers to questions.   Other additions to the
composition routines currently being developed will provide the
following capabilities:

       (1) Construction of network representations of phrases
       for which some constituents are partially or totally
       unspecified.
       (2) Prediction of the composition of the missing
       components in these incomplete phrases.

       One of the most important of our current activities
in semantics is designing and implementing a retrieval system that
will examine. the network structure produced as a result of

parsing, interpret the meaning of the input, and develop and execute a plan for producing an appropriate response. Our short term goal is to respond appropriately to input queries that contain only one verb-like structure and that can be answered from information contained explicitly in the data base. Outputs initially will be YES, NO, or simple noun phrases.

d.   Discourse Analysis and Pragmatics

During the current contract period, we continued to collect and analyze protocols of task-oriented dialogs. Previously, with the cooperation of personnel from the Naval Postgraduate School in Monterey, California, we had conducted experiments for the data management task domain in which naval officers queried specifications and performance characteristics of submarines in the U.S., Soviet, and British fleets. Also, in conjunction with the computer based consultant project at SRI, we gathered dialogs from the workstation environment for our second task domain. Currently, with the help of the Naval Electronics Laboratory Center in San Diego, we are recording protocols using a new data management scenario involving U.S. and Soviet ships in the Mediterranean. Further experiments for the computer consultant task are planned.

The protocols already gathered have been analyzed to identify modifications in the syntax and vocabulary, as described in the section on language definition. In addition, they have been examined for instances of ellipsis and anaphoric

reference.   Guided by this analysis, we have designed   and
implemented a preliminary discourse package that handles the
simpler forms of ellipsis and anaphora found in the dialogs.   A
history of previous utterances is kept, and after an utterance is
successfully parsed, references are resolved using the immediately
preceding   utterance   as   context.   In addition, elliptical
utterances are completed, if possible, by comparing them with
parts of the preceding utterance and adapting the structure in
which the corresponding parts are embedded.

           We are in the process   of   augmenting   these
procedures   in   several ways.   The availability of multiple
partitions and contexts in the semantic net will enable us to
identify 'focus spaces', that is, regions that are directly
related to the current discourse.   Use of this mechanism will
limit the portion of the net that has to be considered in
resolving references; it will be particularly helpful for the
computer consultant task domain, which is more structured and
considerably more complicated than the data management one.   Four
steps are entailed in making these extensions:

      (1) Representation of the focus partition   in   the
      semantic network.
      (2) Preparation of a set of criteria for deciding when
      to establish a new focus space and what to put in it.
      (3) Development of a set of heuristics for deciding
      which spaces to search and when to search them in order

to resolve uncertainties of reference.

(4) Integration of the focus space mechanism with the current discourse package.

In the milestone system, we expect to be resolving simple anaphoric references from the discourse context using the focus space structure. Furthermore, resolution will be performed at the phrase level rather than waiting until the structure for a complete utterance has been produced. We also will introduce a preliminary form of prediction on the basis of the discourse routines. The discourse history will be extended to keep track of topics recently talked about, and procedures will be developed to change the priority (score) of related elements in the language definition accordingly.

C.   References

Becker, Richard, and Poza, Fausto.   Acoustic Processing in the SRI Speech Understanding System, IEEE Transactions on Acoustics, Speech and Signal Processing, 1975 [in press].

Deutsch, Barbara G. The Structure of Task-Oriented Dialogs, Contributed   Papers,   IEEE   Symposium on Speech Recognition, Carnegie-Mellon University, Pittsburgh, Pennsylvania, 15-19 April 1974. IEEE, New York, 1974, 250-254.

Hart, Peter E. Progress on a Computer Based Consultant, Technical   Note 99, Artificial Intelligence Center, Stanford

Research Institute, Menlo Park, California, January 1975.

Hendrix, Gary G. Expanding the Utility of Semantic Networks Through Partitioning. Advance Papers of the Fourth International Joint Conference on Artificial Intelligence, Tbilisi, Georgia, USSR, 3-8 September 1975.

Newell, Allen, et al. Speech Understanding Systems. North-Holland Publishing Company, Amsterdam, 1973.

Nilsson, Nils J., et al. Artificial Intelligence--Research and Applications. Annual Report, Project 3805, Artificial Intelligence Center, Stanford Research Institute, Menlo Park, California, May 1975.

Paxton, William H. A Best-First Parser, Contributed Papers, IEEE Symposium on Speech Recognition, Carnegie-Mellon University, Pittsburgh, Pennsylvania, 15-19 April 1974. IEEE, New York, 1974, 218-225. [IEEE Transactions on Acoustics, Speech and Signal Processing, in press.]

Paxton, William H., and Robinson, Ann E. A Parser for a Speech Understanding System. Advance Papers, International Joint Conference on Artificial Intelligence, Stanford, California, 20-23 August 1973. Stanford Research Institute, Menlo Park, California, 1973, 216-222.

Ritea, H. Barry. A Voice-Controlled Data Management System. Contributed Papers, IEEE Symposium on Speech Recognition,

Carnegie-Mellon University, Pittsburgh, Pennsylvania, 15-19 April 1974. IEEE, New York, 1974, 28-31.

Robinson, Jane J. Performance Grammars. Invited Papers, IEEE Symposium on Speech Recognition, Carnegie-Mellon University, Pittsburgh, Pennsylvania, 15-19 April 1974. To be published by Academic Press, New York, 1975.

Walker, Donald E. Speech Understanding Research. Annual Report, Project 1526, Artificial Intelligence Center, Stanford Research Institute, Menlo Park, California, February 1973. (a)

Walker, Donald E. Speech Understanding Through Syntactic and Semantic Analysis. Advance Papers, International Joint Conference on Artificial Intelligence, Stanford, California, 20-23 August 1973. Stanford Research Institute, Menlo Park, California, 1973, 208-215. (b)

Walker, Donald E. Speech Understanding, Computational Linguistics, and Artificial Intelligence. In: Computational and Mathematical Linguistics, Proceedings of the International Conference on Computational Linguistics, Volume I. Edited by Antonio Zampolli. Casa Editrice Leo S. Olschki, Firenze, 1973. (c)

Walker, Donald E. Speech Understanding Research. Annual Report, Project 1526, Artificial Intelligence Center, Stanford Research Institute, Menlo Park, California, May 1974. (a)

Walker, Donald E.   The   SRI   Speech   Understanding   System.
Contributed   Papers,   IEEE   Symposium   on   Speech   Recognition,
Carnegie-Mellon University, Pittsburgh, Pennsylvania, 15-19 April
1974.   IEEE,   New   York,   1974,   32-37.   [IEEE   Transactions   on
Acoustics, Speech and Signal Processing, in press.] (b)

Walker, Donald E., et al.   Speech   Understanding   Research.
Annual   Report,   Project   3804,   Artificial   Intelligence   Center,
Stanford Research Institute, Menlo Park, California, June 1975.

Winograd, Terry. Procedures as a Representation for Data in
a Computer Program for Understanding Natural Language. Report
MAC-TR-84, Project MAC, Massachusetts Institute of Technology,
Cambridge,   Massachusetts,   1971.   [Published   as   Understanding
Natural Language, Academic Press, New York, 1972.]